AN EXPERIMENTAL STUDY OF FOUR PROCEDURES FOR AGGREGATING SUBJECTIVE PROBABILITY ASSESSMENTS

Gregory W. Fischer

Decisions and Designs, Incorporated

Prepared for:

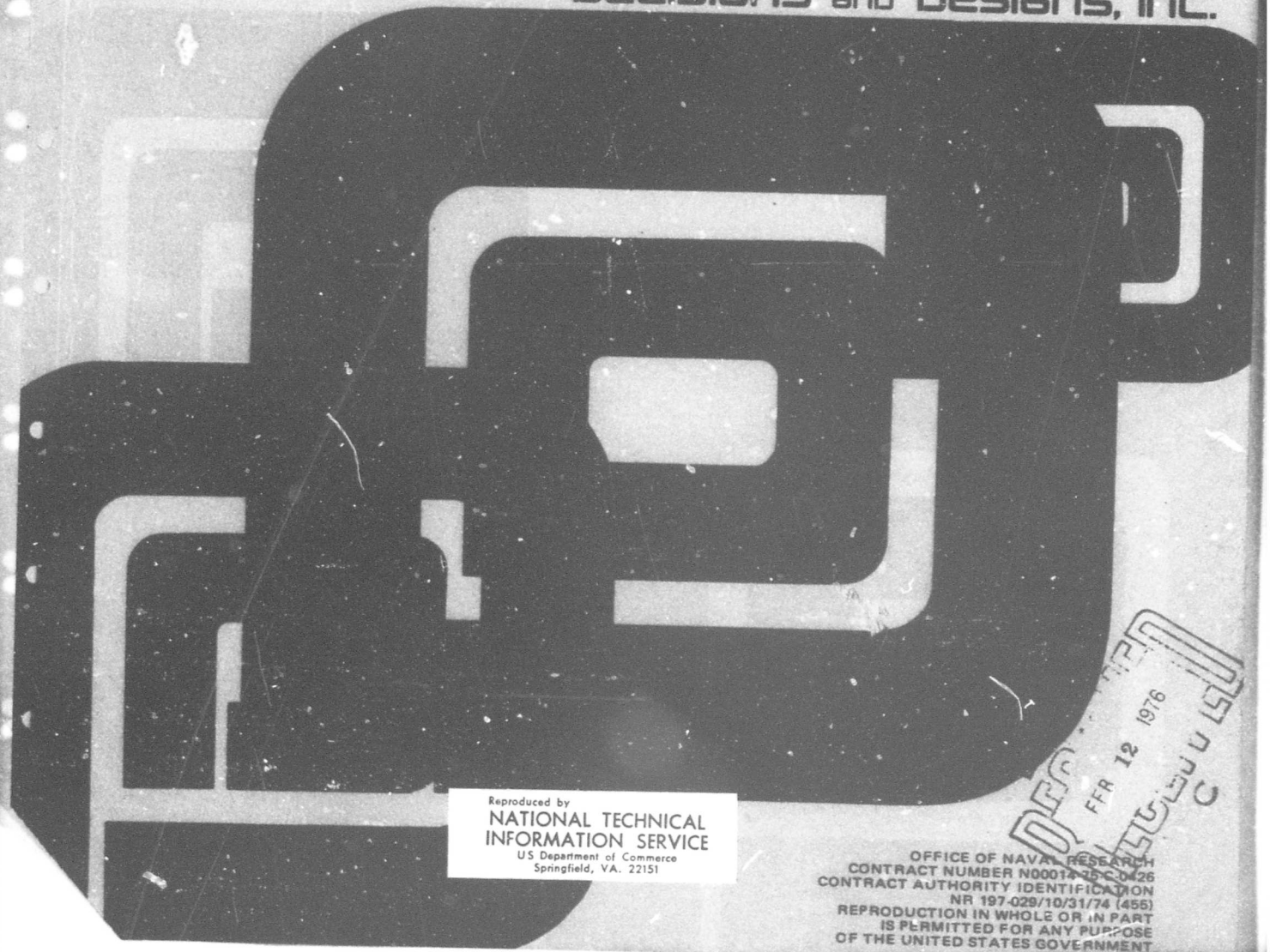
Office of Naval Research

December 1975

An Experimental Study of Four Procedures for Aggregating Subjective Probability Assessments

G.W. Fischer

DECISIONS and DESIGNS, INC.



OFFICE OF NAVAL RESEARCH
CONTRACT NUMBER N00014-75-C-0426
CONTRACT AUTHORITY IDENTIFICATION
NR 197-029/10/31/74 (455)

UNCLASSIFIED

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

REPORT DOCUMENTATION PAGE		READ INSTRUCTIONS BEFORE COMPLETING FORM
1. REPORT NUMBER NR 197-029	2. GOVT ACCESSION NO.	3. RECIPIENT'S CATALOG NUMBER
4. TITLE (and Subtitle) An Experimental Study of Four Procedures for Aggregating Subjective Probability Assessments		5. TYPE OF REPORT & PERIOD COVERED Technical Report 10 Jan 1975-9 Jan 1976 6. PERFORMING ORG. REPORT NUMBER
7. AUTHOR(s) Gregory W. Fischer		8. CONTRACT OR GRANT NUMBER(s) Contract No. N00014-75-C-0426
9. PERFORMING ORGANIZATION NAME AND ADDRESS Decisions and Designs, Inc. 7900 Westpark Drive, Suite 100 McLean, Virginia 22101		10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS Project Element 124209
11. CONTROLLING OFFICE NAME AND ADDRESS Dept. of the Navy, Office of Naval Research 800 N. Quincy Park Arlington, Virginia 22217		12. REPORT DATE December 1975 13. NUMBER OF PAGES 22
14. MONITORING AGENCY NAME & ADDRESS(if different from Controlling Office)		15. SECURITY CLASS. (of this report) UNCLASSIFIED 15a. DECLASSIFICATION/DOWNGRADING SCHEDULE

16. DISTRIBUTION STATEMENT (of this Report)

Approved for public release; distribution unlimited.

17. DISTRIBUTION STATEMENT (of the abstract entered in Block 20, if different from Report)

N/A

18. SUPPLEMENTARY NOTES

N/A

19. KEY WORDS (Continue on reverse side if necessary and identify by block number)

Probability Assessment
Decision Analysis
Group Aggregation Procedures
Statistical Group Estimate
Statistical Averaging
Delphi Technique
Delbecq Method

20. ABSTRACT (Continue on reverse side if necessary and identify by block number)

A number of studies have shown that a consensus probability distribution, obtained by averaging together the assessments of individuals, typically outperforms almost all individuals. The present study evaluated several strategies provided for improving upon this averaging approach.

DD FORM 1 JAN 73 1473 EDITION OF 1 NOV 65 IS OBSOLETE

UNCLASSIFIED

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

TECHNICAL REPORT 75-7

AN EXPERIMENTAL STUDY OF FOUR PROCEDURES FOR AGGREGATING SUBJECTIVE PROBABILITY ASSESSMENTS

by

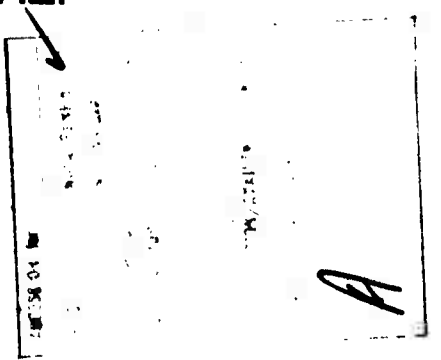
Gregory W. Fischer

for

Department of the Navy
Office of Naval Research
800 North Quincy Street
Arlington, Virginia 22217

December 1975

DECISIONS and DESIGNS, INC.
Suite 100, 7900 Westpark Drive
McLean, Virginia 22101
(703) 821-2828



AN EXPERIMENTAL STUDY OF FOUR PROCEDURES FOR AGGREGATING SUBJECTIVE PROBABILITY ASSESSMENTS

TABLE OF CONTENTS

ABSTRACT

A number of studies have shown that a consensus probability distribution, obtained by averaging together the assessments of individuals, typically outperforms almost all individuals. The present study evaluated several strategies for improving upon this averaging approach. These strategies provide for some type of interjudge interaction.

No between-procedure differences were obtained. In addition, a re-analysis of data from a previous study in which statistically significant between-procedure differences were obtained suggests that these differences were too small to be of practical significance to the applied decision analyst.

Based on these results and a review of the relevant literature, two conclusions emerge: (1) subjective probability distributions can be substantially improved by aggregating the opinions of a group of experts rather than relying on a single expert, and (2) from a practice standpoint, there is no evidence to suggest that the method used to aggregate these opinions will have a substantial effect on the quality of the resulting subjective probability distribution.

AN EXPERIMENTAL STUDY OF FOUR PROCEDURES FOR AGGREGATING SUBJECTIVE PROBABILITY ASSESSMENTS

1.0 INTRODUCTION

The formal analysis of decisions under uncertainty requires three types of subjective inputs. The first, and possibly most important, class of inputs relates to the structure of the problem itself. Someone, usually an important decision maker, must realize that a decision has to be made; a set of alternative courses of action must be generated; and a probability model relating actions to outcomes must be developed. To date, decision-oriented psychologists have generally ignored these stages in the process of structuring a decision problem. The second class of subjective inputs deals with the evaluation of outcomes. Here, goals or objectives must be specified and a quantitative utility model developed. During the past decade, a large amount of formal and psychological research has been addressed to this problem of multi-attribute utility assessment. Finally, subjective probability assessments relating actions to outcomes are typically required for the probability model.

This third class of subjective inputs has attracted by far the most attention from psychologists, with most studies focusing on the subjective probability assessment process of individuals. Formal decision analyses, however, are generally conducted by large organizations which have at their disposal many experts whose opinions may be brought to bear on the assessment of the probabilities of uncertain events. As a consequence, applied decision analysts often find themselves faced with the problem of aggregating the conflicting probability assessments of a group of experts. The present report first summarizes the findings of the limited number of studies addressed to this issue, then presents the results of a recently conducted experiment which compared different group aggregation procedures.

2.0 STUDIES OF GROUP AGGREGATION

Psychologists interested in group versus individual performance have found repeatedly that groups outperform individuals at simple point estimation tasks (Steiner, 1972). In such a task, the subject or group is asked to make a point estimate of an uncertain quantity, such as the length of a line or the area of a rectangle. The early studies in this area suggested that groups typically outperform individuals. The enthusiasm over this apparent superiority of groups was considerably dampened, however, when it was discovered that most of the improvement achieved by groups could be attributed to the well known benefits of statistical averaging. A simple simulation study conducted by Huber and Delbecq (1972) amply illustrates the benefits of statistical averaging in point estimation tasks. In one of their examples they assumed that each expert's opinion was sampled from a normal distribution with mean equal to the true parameter value and a standard deviation equal to 10% of the possible scale range. The expected absolute error for one randomly selected expert is equal to 7.5% of the scale range; taking the mean estimate of five randomly selected experts, it declines to 3.4% of the scale range, and for the mean estimate of ten experts to 2.5% of the scale range.

Similar results have been obtained in experimental studies of point estimation. Dalkey (1969) asked 29 subjects to make point estimates of historical quantities such as the U.S. gross national product in 1965. To evaluate their judgments he used the error score $E = \ln|\hat{\theta}/\theta|$, where $\hat{\theta}$ is the estimated value and θ the true value. He then compared the average error of individual estimates with the average error of all possible "statistical group estimates"[1], for groups ranging in size from 2 to 29. Averaging over groups of five subjects reduced the average error score by 42%; averaging over all 29 subjects reduced it by 65%. Here, as well as in the Huber and Delbecq simulation study, it is clear that the benefits of averaging are subject to diminishing marginal returns. In both examples, most of the reduction in error is achieved by going from one to five judges.

The statistical averaging approach can be applied to subjective probability assessments as well as to point estimates. Suppose, for example, that we want a subjective probability distribution over the uncertain variable $\tilde{x}$, where $\tilde{x}$ may be either continuous or discrete, and that a set of experts have assessed the subjective distribution functions $f_1(\tilde{x}), f_2(\tilde{x}), \ldots, f_n(\tilde{x})$, where $f_i(\tilde{x})$ denotes the distribution function assessed by the

[1] In this paper the term "statistical group estimate" is used to denote an estimate obtained by averaging together the individual estimates of all members of the group.

i-th expert, and where each of the $f_i(\tilde{x})$ satisfies the formal properties of a probability distribution function. Then Winckler (1968) has shown that $g(\tilde{x}) = \sum_{i=1}^{n} w_i f_i(\tilde{x})$ also satisfies the properties of a probability distribution function provided that $\sum_{i=1} w_i = 1$, for $0 \leq w_i \leq 1$. The probability distribution $g(x)$, of course, is simply a weighted average of the $f_i(x)$. In the simple averaging case $w_i = 1/n$.

Several studies have assessed the benefits of statistically averaging probability distributions assessed by different judges. Each of these studies has used proper scoring rules to evaluate the quality of individual versus group average assessments. The primary function of scoring rules is to motivate the assessor to make "honest" assessments. A scoring rule is said to be strictly proper if it satisfies the property that an assessor can maximize his subjectively expected score only if he states his "true beliefs." Two commonly utilized scoring rules which satisfy this property are the logarithmic scoring rule $L(p_1, p_2, \ldots, p_m) = \log p_k$ and the quadratic scoring rule $Q(p_1, p_2, \ldots, p_m) = 2 p_k - \sum_{i=1} p_i^2$, where $(p_1, p_2, \ldots, p_m)$ is the vector of probabilities assigned to the set of events of interest and p_k is the probability assigned to the event which actually occurs. [Stael von Holstein (1970) and Murphy and Winkler (1970) provide a more complete discussion of the mathematical properties of these scoring rules.]

In the studies of concern here, scoring rules were used not only to motivate assessors, but also to evaluate the quality of their assessments. In the first of these, Winkler (1971) conducted a season-long study in which subjects assigned probabilities to various point spreads in Big Ten and National Football League games. He then evaluated these assessments using both the logarithmic and quadratic scoring rules. Because these (and all other proper) scoring rules are convex, it can be shown that the score of the average distribution function must exceed the average individual score. But Winkler found that the average distribution did much better than this, outperforming 95% of the subjects in the study. Using the quadratic scoring rule, the group average function outperformed the average individual score by 5% to 10%; using the more sensitive logarithmic rule, it outperformed the average individual score by 26% to 28%. Stael von Holstein (1971, 1972) obtained similar results in studies of weather forecasting and stock market projections. Together, these three studies strongly suggest that multiple expert opinions should be obtained when possible. For the group average functions typically outperformed all but one or two individuals. Moreover, further analyses carried out by Winkler (1971) suggest that it is almost impossible to determine from past performance which individuals will outperform the group function. Even differential weighting schemes based on past success offer only slight improvement over equal weighting.

Given that statistical averaging improves probability assessments, it seems natural to ask: Can some form of interaction between experts provide benefits over and above those of averaging? A number of authors (Dalkey and Helmer, 1963; Gustafson, et al., 1973) have argued that there are reasons for believing that direct face-to-face interactions might, in some cases, actually generate poorer assessments. This expectation is based in part on results from empirical studies of group problem solving. (See Steiner, 1972, for an integrative review of the literature.) It has been found, for example, that high status group members tend to dominate group decisions even when their proposed solutions are inferior to those of lower status group members (Torrance, 1954). In addition, self-confident assertive members are more likely to get their position adopted (Johnson and Torcivia, 1967), and to dominate the discussion process, thus reducing the input of potentially more knowledgeable members (Dalkey and Helmer, 1963). Other studies have shown that groups sometimes focus on a simple aspect of a problem and come to a decision before all aspects of the problem have been considered.

Two studies suggest that these negative aspects of group discussion processes may more than offset the potential benefits of discussion in probability assessment tasks. Goodman (1970) asked 27 individuals to assess likelihood ratios in a Bayesian inference task. She then assigned 24 of these individuals to four-person groups which reassessed the same set of likelihood ratios. Responses were scored in terms of the accuracy ratio SLLR/BLLR, where SLLR is the subjective log likelihood ratio, and BLLR is the Bayesian log likelihood ratio. An accuracy ratio is said to be conservative if SLLR/BLLR < 1. For four of the six groups, the group consensus SLLRs were significantly more conservative (and further from optimal) than the average of the pre-group SLLRs assessed by the group members. But for the other two groups, the group SLLRs were significantly less conservative (and closer to optimal). Moskowitz (1971) used a similar design, asking subjects to estimate default probabilities for hypothetical loan applicants based on three independent sources of data of known diagnosticity. Using accuracy ratios to measure performance, he found that statistical groups substantially outperformed real groups, with mean accuracy ratios of .90 and .63, respectively. The superiority of the statistical groups was greatest for data with high diagnosticity. The real groups, in fact, performed better on items involving data of low diagnosticity. These two studies, then, provide limited support for the argument that statistical averaging is preferable to direct interaction for probability assessments.

Next, we will consider two approaches designed to realize the benefits of direct interaction without incurring its costs. The first of these approaches, the Delphi technique (Dalkey and Helmer, 1963) is older and considerably better

known. The Delphi technique relies on successive iteration in which judges make anonymous assessments and are then given anonymous statistical feedback about the assessments of the other judges. In informationally richer variants of the Delphi procedure, judges give written explanations of their responses. In order to ensure anonymity, these explanations are carefully edited before being distributed. At the end of the final iteration, the individual estimates are averaged together to provide the group estimate. Proponents of the Delphi approach argue that by preserving anonymity, it overcomes the liabilities of face-to-face groups, and that by providing feedback it should be superior to statistical groups. In addition, it shares with statistical groups the practical merit of not requiring that the judges physically be brought together. On the negative side, the iteration process may be quite time consuming, particularly if editing of written explanations is required.

An alternative approach, advocated by Andre Delbecq and his colleagues at the University of Wisconsin, involves four steps. First, each judge makes his own initial estimate. Next, each judge displays his initial opinion to the group, thus assuring that all opinions are at least presented. Then, the group members discuss their estimates and the reasoning behind them. Finally, each judge anonymously makes his final estimates. These final estimates are then averaged together to determine the final group consensus opinion. This procedure, which we will term the Delbecq method, differs from the Delphi method in only one important respect: Direct discussion is substituted for statistical feedback. Clearly, only experimentation can determine whether either or both of these approaches is superior to simple statistical averaging.

To date, only one study has compared the three approaches. Gustafson, et al., (1973) asked groups of subjects to make inferences about the gender of randomly selected college students based on only a single datum, either the height or weight of the student in question. Four types of groups were studied. In addition to the Delphi, Delbecq, and simple averaging procedures, a fourth condition was included in which subjects first talked the problem over, then made anonymous estimates which were then averaged together. This talk-estimate procedure differed from the Delbecq method only in that group members did not make prior estimates to be shown to the group. All responses were recorded on log odds scales, and scored using

$$E = 100 \left| \frac{\log \Omega_{\cdot\cdot} - \log \Omega}{\log \Omega} \right|$$

where Ω denotes the actuarially correct odds, $\hat{\Omega}$ the group estimated odds, and logarithms are to the base 10. An analysis of variance of these error scores indicated a highly significant treatment effect. The average error scores for the Delbecq groups ($\bar{E}$ = 78%) were considerably lower than those of the single averaging groups ($\bar{E}$ = 114%), the talk-estimate groups ($\bar{E}$ = 111%), and the Delphi groups ($\bar{E}$ = 128%). (In defense of the Delphi procedure, it should be noted that only one iteration was carried out, and only anonymous feedback on the actual estimates of the other subjects was provided.) Although it is hazardous to generalize from a single study, these results clearly suggest that the Delbecq procedure may provide a superior means of aggregating the opinions of a group of probability assessors.

3.0 GROUP PREDICTIONS OF SUCCESS IN COLLEGE

3.1 General

The primary goal of the study to be reported here was to determine the extent to which the results obtained by Gustafson, et al., would generalize to other types of probability assessment tasks. The primary change in the design was to substitute a true group consensus condition for the talk-estimate condition. The other deviations from the Gustafson, et al., design were as follows:

a. The uncertain event of interest had four outcome classes instead of two.

b. Subjects responded on a probability rather than a log odds scale.

c. Subjects were motivated by a pay-off system based on a truncated logarithmic scoring rule.

d. Subjects were given trial-by-trial feedback concerning both the true event and their score for the trial.

3.2 Subjects

Two types of subjects participated. Those from the introductory psychology subject pool received one hour of experimental credit plus whatever incentive pay they earned. All other subjects earned $1.50 plus whatever incentive pay they earned. All subjects were Duke University students. Eight groups of three subjects served in each of the four experimental conditions.

3.3 Task

Subjects were asked to make predictions about the freshman grade point average (GPA) of 10 randomly selected students from a recent Duke freshman class. Each case description, or profile, contained the following pieces of information: gender, high school GPA, SAT math score, and SAT verbal score. Based on this information, subjects were asked to assess the probability that the freshman GPA of the student described fell into the ranges: 0-2.49, 2.50-2.99, 3.00-3.49, and 3.50-4.00. Because these intervals are mutually exclusive and exhaustive, subjects were instructed to make sure that the set of probabilities summed to 1.00. Of the 800 sets of probabilities assessed, nine failed to satisfy this criterion. These nine sets of estimates were normalized to sum to 1.00.

3.4 Scoring Rule

The incentive pay-offs were based on the truncated logarithmic scoring rule

$$S = \begin{cases} -500 \text{ if } P_t = 0 \\ 50\,[2 + \log_{10}(P_t)] \text{ if } 0 < P_t \leq 1.00 \end{cases}$$

where P_t is the probability assigned to the interval in which the students' grade point average actually fell.[1] These scores were then transformed into monetary amounts using the exchange rate 5 points per penny. Thus, on each trial, a subject could win up to 20¢ or lose as much as $1. As can be seen from Table 1, this scoring rule imposes heavy sanctions for the assignment of very small probabilities to the true event. It becomes, however, quite insensitive to differences between assessments above .40, and very insensitive to differences between assessments above .75.

To simplify their task, subjects were asked to assign probabilities which were multiples of .05, except in the regions 0 to .05 and .95 to 1.0, where multiples of .01 were permitted. After each trial, subjects were given outcome feedback and asked to record their score for the trial.

In the statistical groups condition, each subject received incentive pay based on his own assessments. In the talk-to-consensus condition, incentive pay was based on the collective consensus assessment, thus providing subjects with an incentive to actively participate. To provide a similar incentive in the Delphi and Delbecq groups, each subject's pay was based on the average score of the final estimates of all three members of the group.

3.5 Results

To provide a benchmark for evaluating the scores received by groups in the various experimental conditions, it is useful to consider the naive strategy of ignoring the information provided and simply assigning a probability of .25 to each of the four GPA ranges. This strategy is quite reasonable, in fact, because the four GPA ranges were almost equally likely, with marginal probabilities ranging from .18 to .30. As can be seen from Table 1, this strategy assures a score of 70 on each trial, and a total score of 700 over a 10-trial session.[2]

---

[1] The truncated logarithmic rule is not strictly proper. But for practical purposes, the fines associated with a negative infinity score cannot be credibly threatened.

[2] Using the 5-points-per-penny exchange rate, a subject would thus earn $1.40 in incentive pay. This naive strategy was, in fact, used to establish the points-to-pennies exchange rate.

TABLE 1

THE TRUNCATED LOGARITHMIC SCORING RULE[1]

Probability Assigned to True Event	Score
0	-500
.01	0
.02	15
.03	24
.04	30
.05	35
.10	50
.15	59
.20	65
.25	70
.30	74
.35	77
.40	80
.45	83
.50	85
.55	87
.60	89
.65	91
.70	92
.75	94
.80	95
.85	96
.90	98
.95	99
.96	99
.97	99
.98	100
.99	100
1.00	100

[1]Rounded to the nearest integer.

Of the 24 subjects who served in the statistical groups condition, 20 outperformed the naive equiprobable assessment strategy. The median score for these 24 subjects was 727, but the mean was only 682. Two individuals assigned a 0 probability to a true event, thus receiving a very low total score and pulling the overall mean down accordingly.

The principal findings of the study are summarized in Figure 1. As might be expected, all four opinion aggregation procedures substantially outperformed both the naive equiprobability assessment strategy and the average individual subject. What is most striking, however, is the virtual equality of the mean scores for the four aggregation procedures. A simple one-way analysis of variance produced an F-statistic of .45 ($p \leq .999$), with the treatment effects explaining only 4.6% of the total variance. While one can never confirm the null hypothesis of no treatment effects, the present data are certainly consistent with it. Differences of the magnitude observed here, even if statistically significant, would clearly be of no practical interest to the applied decision analysis.

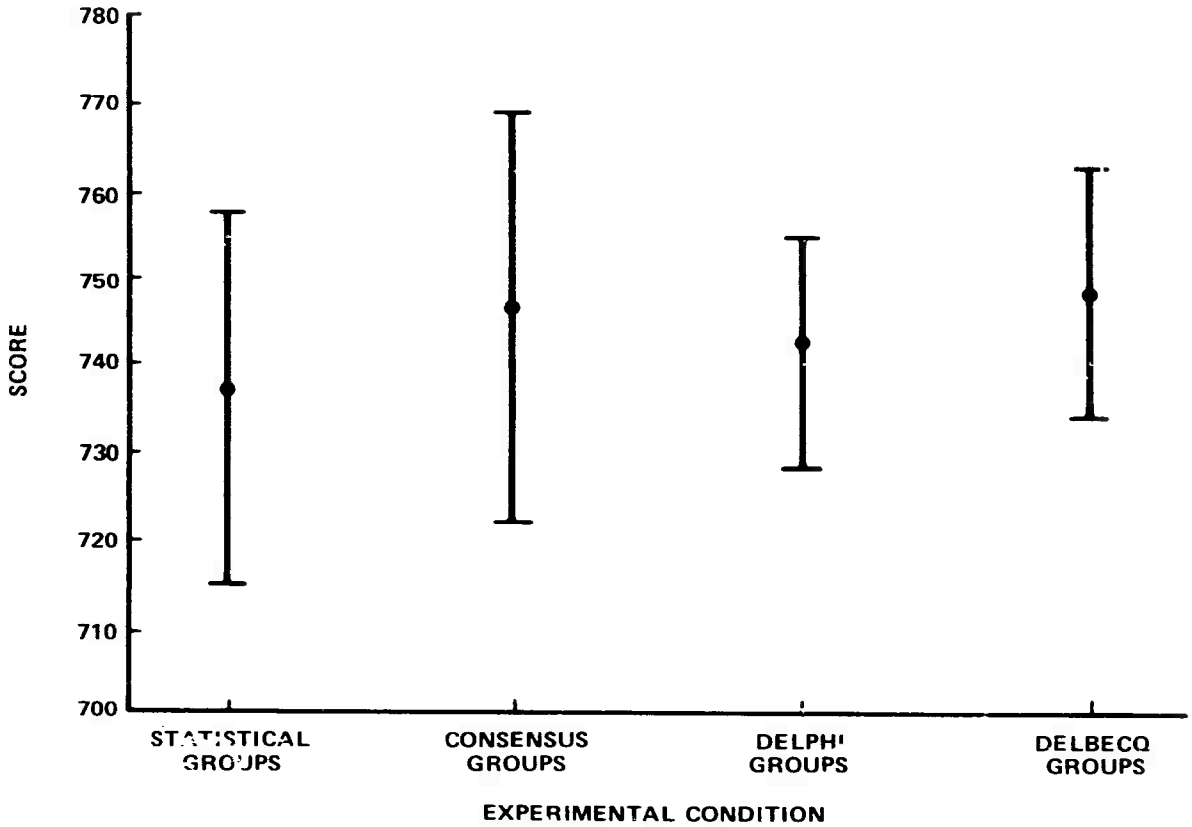


Figure 1

MEANS AND 95% CONFIDENCE INTERVALS
FOR THE AVERAGE SCORE IN EACH CONDITION

4.0 DISCUSSION AND CONCLUSIONS

How can we account for the apparent discrepancy between the results obtained by Gustafson, et al., (1973) and those reported here? One obvious criticism of the present study is that there were only eight groups in each experimental condition, thus producing a fairly high probability of a Type II error. This argument is at least partially offset, however, by the fact each data point should be quite stable. For the total score for each group is obtained by summing over 10 items. And, as noted above, the confidence intervals suggest that any between-procedures differences which might exist are not of sufficient magnitude to be of any practical interest.

Comparing the studies, a number of procedural variations might explain the discrepant findings. First, the prediction tasks themselves differed. Even lacking a good actuarial model of the prediction task used in the present study, it seems likely that the Gustafson, et al., task provided subjects with information of relatively high diagnosticity as compared to the present task. Second, subjects in the Gustafson, et al., study responded on a log odds scale in contrast to the simple probability scale used in the present study. Third, subjects in the present study were motivated by scoring rule feedback. Perhaps procedural variations are of little importance when subjects are highly motivated and provided with feedback on the quality of their estimates. Finally, the two studies used different dependent variables. As will be shown below, the Gustafson, et al., error score is highly sensitive to certain types of small differences. While this may or may not be related to the statistical significance of their findings, it clearly casts doubt on the substantive significance of their findings.

To illustrate this point, we will consider Gustafson, et al., item #1 which produced the largest difference between the Delphi and Delbecq procedures. In particular, this item stated that the height of a college-age Midwesterner was 68 inches. Actuarially, this fact supports the male hypothesis, with a likelihood ratio of 1.8. Assuming males and females to be equally likely to be sampled, this yields an actuarial posterior odds of 1.8. Based on Gustafson, et al., Table 1 and Figure 2, the average odds assessment for the Delbecq groups was approximately 9.55, and for the Delphi groups, approximately 19.05[1]. Clearly, both types of groups overshot the actuarial odds, but the Delphi groups were worse. Using the percent error score

[1]These reconstructions are only approximate. But the conclusions are not sensitive to minor errors.

$$E = 100\left|\frac{\log\hat{\Omega} - \log\Omega}{\log\Omega}\right|$$

produces a score of 285 for the Delbecq groups and 401 for the Delphi groups. Since low scores are good, the Delbecq groups appear to be substantially better on this item, which, it should be repeated, produced the largest between-groups difference. Converting from log odds to probabilities, however, we find that this apparently large effect is, in fact, quite trivial. The actuarial posterior probability is .64; the mean posterior probability for the Delbecq groups was roughly .91; and the mean posterior probability for the Delphi groups was .95. Judged in this light, both estimates are far from optimum, and the difference between the two estimates is quite small. Therefore, had absolute deviations from the actuarial probabilities been used as the dependent variable in the Gustafson, et al., study, the between-procedure differences would have appeared much less substantial. Some rough calculations suggest that, over all eight items, the Delbecq method produced probabilities which were, on the average, .04 closer to the optimal probabilities.

Which dependent variable provides a more appropriate measure of performance? In decision analysis, subjective likelihood judgments are used in the computation of expected utilities. Since expected utility calculations are linear in probability, not odds or log odds, it seems reasonable to argue that deviations from optimal inference should be measured in terms of absolute deviations from Bayesian probabilities. Given the general insensitivity of linear models to small parameter changes [see von Winterfeldt and Edwards (1973)], a difference of .04 in the probability assigned to an event seems unlikely to have a substantial impact on expected utility calculations. Thus, the differences obtained by Gustafson, et al., appear to be too small to be of practical interest.

Similar criticisms may be directed toward studies using accuracy ratios as the dependent variable. Suppose, for example, that the Bayesian odds are 999:1 and the subjectively assessed odds are 99:1. This yields an accuracy ratio of .665, which suggests a very substantial deviation from optimality. In terms of probabilities, however, the subjectively assessed odds imply a probability of .99 which differs only trivially from the Bayesian probability of .999. Again, choice of an inappropriate dependent variable may make small effects (or deviations from optimality) look like large ones.

To summarize, the present study fails to replicate the findings of Gustafson, et al., (1973). Given the multiple design differences, it is impossible to determine the cause of this discrepancy. In addition, it has been argued that the dependent variable used by Gustafson, et al., was misleading. Based on a rough reconstruction of their data, it appears that the effects they obtained were too small to be of practical significance.

From an applied standpoint, then, there seems to be little reason to prefer one group aggregation procedure over another. The existing data strongly suggest, however, that many experts are better than one. Whenever confronted by an important uncertain quantity, decision analysts would be well advised to seek the opinions of at least three to five experts.

5.0 REFERENCES

Dalkey, N.C. The Delphi method: An experimental study of group opinion. RAND Memorandum Rm-5888-PR, June, 1969.

Dalkey, N.C., and Helmer, O. An experimental application of the Delphi method to the use of experts. Management Science, 1963, 9.

Goodman, B.C. Risky decisions by individuals and groups. Doctoral dissertation, Department of Psychology, The University of Michigan, 1970.

Gustafson, D.H., et al. A comparative study of differences in subjective likelihood estimates made by individuals, interacting groups, Delphi groups, and nominal groups. Organizational Behavior and Human Performance, 1973, 9, 280-291.

Huber, G.P. and Delbecq, A. Guidelines for combining the judgments of individuals in decision conferences. Academy of Management Journal, 1972, 161-174.

Johnson, H.H., and Torciria, J.M. Group and individual performance on a single stage task as a function of distribution of individual performance. Journal of Experimental Social Psychology 1967, 3, 266-273.

Moskowitz, H. Conservatism in group information processing behavior under varying management information systems. Krannert School of Industrial Administration, Purdue University, Paper No. 333, 1971.

Murphy, A.H., and Winkler, R.L. Scoring rules in probability assessment and evaluation. Acta Psychologica, 1970, 34, 273-286.

Stael von Holstein, C.A.S. Measurement of subjective probability. Acta Psychologica, 1970, 34, 146-159.

Stael von Holstein, C.A.S. An experiment in probabilistic weather forecasting. Journal of Applied Meteorology, 1971, 10, 635-645.

Stael von Holstein, C.A.S. Probabilistic forecasting: An experiment related to the stock market. Organizational Behavior and Human Performance, 1972, 8, 139-158.

Steiner, I.D. Group Process and Productivity, New York: Academic Press, 1972.

Torrance, E.P. Some consequences of power differences on decision making in permanent and temporary three-man groups. Research Studies, State College of Washington, 1954, 22, 130-140.

von Winterfeldt, D., and Edwards, W. Costs and payoffs in perceptual research. The University of Michigan, Engineering Psychology Laboratory Technical Report 011313-1-T, 1973.

Winkler, R.L. The consensus of subjective probability distributions. Management Science, 1968, 15, B61-75.

Winkler, R.L. Probabilistic prediction: Some experimental results. Journal of the American Statistical Association, 1971, 66, 675-685.